UDC 004.94
doi:10.31799/1684-8853-2021-2-60-71

Articles

Development and application of problem-oriented digital twins for magnetic observatories and variation stations

A. V. Vorobev[a,b], *PhD, Tech., Associate Professor, orcid.org/0000-0002-9680-5609, geomagnet@list.ru*
V. A. Pilipenko[b,c], *Dr. Sc., Phys.-Math., Principal Researcher, orcid.org/0000-0003-3056-7465*
G. R. Vorobeva[a], *PhD, Tech., Associate Professor, orcid.org/0000-0001-7878-9724*
O. I. Khristodulo[a], *Dr. Sc., Tech., Associate Professor, orcid.org/0000-0002-3987-6582*
[a]*Ufa State Aviation Technical University, 12, K. Marx St., 450008, Ufa, Russian Federation*
[b]*Geophysical Center of the RAS, 3, Molodezhnaya St., 119296, Moscow, Russian Federation*
[c]*Institute of Physics of the Earth of the RAS, 10, b. 1, B. Gruzinskaya St., 123995, Moscow, Russian Federation*

Introduction: *Magnetic stations are one of the main tools for observing the geomagnetic field. However, gaps and anomalies in time series of geomagnetic data, which often exceed 30% of the number of recorded values, negatively affect the effectiveness of the implemented approach and complicate the application of mathematical tools which require that the information signal is continuous. Besides, the missing values add extra uncertainty in computer simulation of dynamic spatial distribution of geomagnetic variations and related parameters.* ***Purpose:*** *To develop a methodology for improving the efficiency of technical means for observing the geomagnetic field.* ***Method:*** *Creation of problem-oriented digital twins of magnetic stations, and their integration into the collection and preprocessing of geomagnetic data, in order to simulate the functioning of their physical prototypes with a certain accuracy.* ***Results:*** *Using Kilpisjärvi magnetic station (Finland) as an example, it is shown that the use of digital twins, whose information environment is made up of geomagnetic data from adjacent stations, can provide the opportunity for reconstruction (retrospective forecast) of geomagnetic variation parameters with a mean square error in the auroral zone of up to 11.5 nT. The integration of problem-oriented digital twins of magnetic stations into the processes of collecting and registering geomagnetic data can provide automatic identification and replacement of missing and abnormal values, increasing, due to the redundancy effect, the fault tolerance of the magnetic station as a data source object. For example, the digital twin of Kilpisjärvi station recovers 99.55% of annual information, and 86.73% of it has an error not exceeding 12 nT.* ***Discussion:*** *Due to the spatial anisotropy of geomagnetic field parameters, the error at the digital twin output will be different in each specific case, depending on the geographic location of the magnetic station, as well as on the number of the surrounding magnetic stations and the distance to them. However, this problem can be minimized by integrating geomagnetic data from satellites into the information environment of the digital twin.* ***Practical relevance:*** *The proposed methodology provides the opportunity for automated diagnostics of time series of geomagnetic data for outliers and anomalies, as well as restoration of missing values and identification of small-scale disturbances.*

Keywords *— digital twins, time series reconstruction, statistical analysis, geomagnetic data, magnetic stations.*

For citation: Vorobev A. V., Pilipenko V. A., Vorobeva G. R., Khristodulo O. I. Development and application of problem-oriented digital twins for magnetic observatories and variation stations. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2021, no. 2, pp. 60–71. doi:10.31799/1684-8853-2021-2-60-71

Introduction

Today, magnetic observatories and variation stations are among the main instruments for observing the geomagnetic field (GMF) and its variations. There are more than 300 ground magnetic stations that record the parameters of the GMF in real time mode. Usually, these magnetic stations are integrated into networks, which for the data consumers are represented as the specialized web-services that provide access to geomagnetic data and have the functionality necessary for their search, preview and download. By the end of 2020, more than 20 such networks of magnetic stations are known, the largest of which are INTERMAGNET, IMAGE, CARISMA, MACCS, MAGDAS, etc.

Outliers, gaps in time series, noise and other anomalies are widespread and still not having a final solution to the problem on the way of processing the received geophysical information. Even for magnetic observatories of the INTERMAGNET network [1, 2], which maintains the highest quality standard, the lengths of the missing fragments occupy a fairly wide range and vary both in time and from station to station. For example, in 2015 the quantity of missing values for station AlmaAta was 36% of the annual operating time, for station Dalat it was more than 12%, for station Sodankyla it was 0.4%, etc. [3].

Multiple anomalies in time series (occurring as a result of measurement errors, registration or noisy information signal), in addition to negatively affecting the efficiency of the implemented

approach to monitoring GMF, also complicate the use of software elements that require compliance with the condition of information signal continuity (calculation of the derivative, Fourier transform, wavelet transform, etc.). In addition, the missing values complicate the problems of computer modeling of the dynamics of the spatial distribution of GMF variations [4, 5] and associated high-level experimental information (indices of geomagnetic activity, perturbation maps, magnetic keograms, etc.) [6].

Until recently, the reconstruction of the GMF observations results was provided by using a linear interpolation or a cubic spline, which is generally acceptable to recover the single gaps, but absolutely unsuitable for imputing long-term fragments. Today more complex approaches to the reconstruction of this type of time series are known. They are based mainly on analytical processing of data in the vicinity of missing fragments, analysis of periodic and seasonal components, as well as the study of Fourier and wavelet spectra of the information signal [7–11]. Usually all of them can be used to reconstruct the missing fragments, which size does not exceed several tens of minutes. The methods provide a methodological error within 15%, require significant computing power, direct human participation and, as a result, are not applicable to large amounts of data. Thus, the existing practice of collecting and registering geomagnetic data using ground magnetic stations is connected with a number of difficulties and limitations, which largely impede the effective conduct of geophysical/heliogeophysical research.

A promising approach to solving the problem can be the creation and integration the problem-oriented digital twins (DT) of magnetic stations into the process of collecting geomagnetic data. The DT allow with a certain accuracy (at the data consumer level) to simulate the work of their physical prototypes [12, 13]. The implementation and development of the proposed concept can significantly increase the efficiency of the operation of separate magnetic stations, as well as reduce the labor intensity of preliminary processing of geomagnetic data.

Analysis of gaps in time series of geomagnetic data and assessment of reliability indicators of ground magnetic stations

An experimental set is provided by the minute data of the IMAGE magnetometer network (https://space.fmi.fi/image/) [14] for 2015, that is the period corresponding to the maximum activity of the 24th solar cycle (January 2009–May 2020).

Table 1 describes the results of assessing the completeness of the time series of 36 stations, where the appearance of a missing value is regarded as a failure of a technical object, i. e., its transition to an inoperative state (State Standard 27.002-2015). Hence, the total idle time TF of the station, corresponding to the number of missing values in the time series, is determined as follows:

$$T_F = T - T_W, \quad (1)$$

where T is an operating time; T_W is a number of informative values (total uptime) for a time period T.

The average time to recover the operating state (equivalent to the mathematical expectation of the missing fragment size) and the average time to failure of the system (equivalent to the average size of the fragment without gaps) can be determined from next expressions:

$$\langle T2R \rangle = \frac{1}{N_F} \sum_{i=1}^{N_F} T2R_i = \frac{T_F}{N_F}; \quad (2)$$

$$\langle T2F \rangle = \frac{1}{N_W + k} \sum_{i=1}^{N_W + k} T2F_i = \frac{T_W}{N_W + k}, \quad (3)$$

where $T2R_i$ and $T2F_i$ are the time until the i-th system recovery after a failure and the time before the i-th system failure, respectively; N_F and N_W are the number of system failures and the number of failover recoveries, respectively; $k = 1$ or $k = 0$, if at the moment of observation beginning the system was in a working or inoperative state, respectively.

The analysis of gaps in the IMAGE network time series demonstrated that in 50% of magnetic stations the expected value of the missing fragment size exceeds 58.5 min. The averaged (over all stations) non-operational time is 1066 min/year. The expected value of the number of failures with recovery for all stations exceeds 45 per year. At the same time, 50% of stations experience more than 17 failures per year. In extreme cases, the total volume of missing fragments of one station can exceed 11.2% (more than 41 days) of the total size of the annual sample, while the average recovery time can reach 10 days or more.

The results indicate that the application of well-known approaches to the reconstruction of time series (linear interpolation, interpolation by cubic splines, as well as the methods described in [7–11]) for most fragments of the missing values of the sources considered here (mainly due to the size missing fragment) is ineffective. In addition, if we are talking about large amounts of information (the results of observing the parameters of the GMF

■ ***Table 1***. Assessment of reliability indicators of magnetic stations of the IMAGE network

IAGA code	Coordinates, degr.				T_W		T_F		N_F	<T2R>, min	<T2F>, min
	GEO		CGM								
	LAT	LON	LAT	LON	min	%	min	%			
NAL	78.92	11.95	76.57	109.96	509551	96.947	16049	3.053	20	802.45	25477.55
LYR	78.20	15.82	75.64	111.03	506314	96.331	19286	3.669	11	1753.27	46028.55
HOR	77.00	15.60	74.52	108.72	466554	88.766	59046	11.234	4	14761.5	116638.5
HOP	76.51	25.01	73.53	114.59	492524	93.707	33076	6.293	49	675.02	10051.51
BJN	74.50	19.20	71.89	107.71	525523	99.985	77	0.015	7	11	75074.71
NOR	71.09	25.79	68.19	109.28	519087	98.761	6513	1.239	144	45.23	3604.77
SOR	70.54	22.22	67.80	106.04	523740	99.646	1860	0.354	43	43.26	12180.0
KEV	69.76	27.01	66.82	109.22	525569	99.994	31	0.006	11	2.82	47779.0
TRO	69.66	18.94	67.07	102.77	524713	99.831	887	0.169	15	59.13	34980.87
MAS	69.46	23.70	66.65	106.36	524144	99.723	1456	0.277	73	19.95	7180.05
AND	69.30	16.03	66.86	100.22	525284	99.94	316	0.06	6	52.67	87547.33
KIL	69.06	20.77	66.37	103.75	523732	99.645	1868	0.355	33	56.61	15870.67
IVA	68.56	27.29	65.60	108.61	486940	92.645	38660	7.355	6	6443.33	81156.67
ABK	68.35	18.82	65.74	101.70	525600	100	0	0	0	–	–
MUO	68.02	23.53	65.19	105.23	492390	93.682	33210	6.318	359	92.51	1371.56
KIR	67.84	20.42	65.14	102.62	525577	99.996	23	0.004	13	1.77	40429.0
SOD	67.37	26.63	64.41	107.33	524905	99.868	695	0.132	12	57.92	43742.08
PEL	66.90	24.08	64.03	104.97	491992	93.606	33608	6.394	8	4201.0	61499.0
JCK	66.40	16.98	63.82	98.94	516366	98.243	9234	1.757	36	256.5	14343.5
DON	66.11	12.50	63.75	95.19	511710	97.357	13890	2.643	19	731.05	26932.11
RAN	65.90	26.41	62.92	106.30	519118	98.767	6482	1.233	130	49.86	3993.22
RVK	64.94	10.98	62.61	93.27	513440	97.686	12160	2.314	61	199.34	8417.05
LYC	64.61	18.75	61.87	99.33	525600	100	0	0	0	–	–
OUJ	64.52	27.23	61.47	106.27	525304	99.944	296	0.056	11	26.91	47754.91
MEK	62.77	30.97	59.57	108.66	511795	97.373	13805	2.627	23	600.22	22251.96
HAN	62.25	26.60	59.12	104.72	520619	99.052	4981	0.948	381	13.07	1366.45
DOB	62.07	9.11	59.64	90.19	524128	99.72	1472	0.28	19	77.47	27585.68
SOL	61.08	4.84	58.82	86.25	512471	97.502	13129	2.498	31	423.52	16531.32
NUR	60.50	24.65	57.32	102.35	525540	99.989	60	0.011	2	30.0	262770.0
UPS	59.90	17.35	56.88	95.95	525600	100	0	0	0	–	–
KAR	59.21	5.24	56.70	85.69	524637	99.817	963	0.183	41	23.49	12796.02
TAR	58.26	26.46	54.88	103.11	525137	99.912	463	0.088	12	38.58	43761.42
BRZ	56.17	24.86	52.66	100.97	523584	99.616	2016	0.384	3	672.0	174528.0
SUW	54.01	23.18	50.21	98.95	487904	92.828	37696	7.172	20	1884.8	24395.2
WNG	53.74	9.07	50.15	86.75	525577	99.996	23	0.004	19	1.21	27661.95
NGK	52.07	12.68	48.03	89.28	525600	100	0	0	0	–	–

Note: GEO is a geographic coordinate system; CGM (Corrected GeoMagnetic) is a geomagnetic coordinate system; the magnetic stations of the auroral cluster are highlighted in gray.

for one year or more), then the application of methods, in the algorithms of which the participation of a person is provided, also becomes very complicated.

Synthesis, modification and validation of digital twin models

The physical prototype of DT is considered as a magnetometric module that registers the northern component (*X*-component) of the GMF vector at the Kilpisjärvi (KIL) station. The research here is considered with spatial clustering of the entire set of magnetic stations in order to identify the reference data sources for subsequent modeling of the parameter.

Assessment of the spatial homogeneity of geographic objects based on the Moran's index for geographic proximity according to the metric [15] revealed between a number of stations located in the range of 66–71° N (see Table 1), the presence of a positive spatial autocorrelation, which indicates that these stations belong to the same spatial cluster with KIL (hereinafter referred to as the "auroral cluster").

A comparative analysis of the correlations of the northern (*X*) component of the geomagnetic disturbance vector of the KIL station with similar parameters of other stations of the auroral cluster (Table 2), as well as a number of additional studies [16, 17] confirmed the validity of the assumption and indicate the possibility of using these data as predicates (features) for modeling the parameter X_{KIL}.

Estimation of the coefficient of determination ($R^2 = 0.999$) demonstrated that for the problem being solved, the approach based on the method of multiple linear regression is the best. Linear regression equation that allows to restore the value of the desired parameter $f(x, \beta)$ from the known values $x_1, \ldots, x_k$ has the form:

$$f(x, \beta) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k = \sum_{j=1}^{k} \beta_j x_j = \mathbf{x}^{\mathrm{T}} \beta, \quad (4)$$

where $\mathbf{x}^{\mathrm{T}} = (x_1, x_2, \ldots, x_k)$ is a vector of regressors; $\hat{\boldsymbol{\beta}} = (\beta_1, \beta_2, \ldots, \beta_k)^{\mathrm{T}}$ is a vector column of coefficients; k is a number of model features.

Taking into account the data in Table 2, it is possible to define the expression (4) as follows:

$$\begin{aligned} X^*_{\text{KIL}} = \alpha + \beta_1 X_{\text{NOR}} + \beta_2 X_{\text{NOR}} + \beta_3 X_{\text{NOR}} + \\ + \beta_4 X_{\text{NOR}} + \beta_5 X_{\text{MAS}} + \beta_6 X_{\text{AND}} + \beta_7 X_{\text{IVA}} + \\ + \beta_8 X_{\text{ABK}} + \beta_9 X_{\text{MUO}} + \beta_{10} X_{\text{KIR}} + \\ + \beta_{11} X_{\text{SOD}} + \beta_{12} X_{\text{PEL}} + \beta_{13} X_{\text{JCK}} + \beta_{14} X_{\text{DON}}, \quad (5) \end{aligned}$$

where $\alpha = 418$ nT is an ordinate offset; β_1, β_2, ..., β_{14} are the coefficients calculated by the least squares method: $\beta_1 = -0.0511992$; $\beta_2 = -0.0791793$; $\beta_3 = 0.011932$; $\beta_4 = 0.5858979$; $\beta_5 = -0.2199333$; $\beta_6 = -0.203925$; $\beta_7 = 0.1138129$; $\beta_8 = 0.6873423$; $\beta_9 = 0.0020214$; $\beta_{10} = -0.2845333$; $\beta_{11} = 0.0170759$; $\beta_{12} = 0.0152406$; $\beta_{13} = 0.0037965$; $\beta_{14} = -0.0263773$.

Mean squared error (MSE) of model (5), which is calculated using the cross-validation procedure, was 11.5 nT. This MSE corresponds to 0.51% of the range of X_{KIL} parameter values for 2015. Pearson's correlation coefficient ($r = 0.999$) and the results of Student's *t*-test (statistical criterion ≈ 0, *p*-value ≈ 1) indicate that the original (X_{KIL}) and synthesized (X^*_{KIL}) data are statistically indistinguishable and belong to the same sample. However, the probability of failure-free operation of model (5) is limited by the probability of failure of at least one of the stations included in the auroral cluster (see Table 1) and, according to the available data, is 77.4%.

■ ***Table 2***. Correlations between X_{KIL} and a similar parameter of other stations

Magnetic stations included in the auroral cluster				Magnetic stations not included in the auroral cluster					
Code	*r*	Code	*r*	Code	*r*	Code	*r*	Code	*r*
NOR	0.872	ABK	0.986	NAL	−0.164	LYC	0.642	UPS	0.218
SOR	0.933	MUO	0.957	LYR	−0.129	OUJ	0.617	KAR	0.142
KEV	0.978	KIR	0.958	HOR	0.015	MEK	0.432	TAR	0.176
TRO	0.985	SOD	0.909	HOP	0.015	HAN	0.384	BRZ	0.098
MAS	0.99	PEL	0.875	BJN	0.427	DOB	0.363	SUW	−0.045
AND	0.987	JCK	0.845	RAN	0.053	SOL	0.262	WNG	−0.017
IVA	0.975	DON	0.820	RVK	0.694	NUR	0.274	NGK	−0.044

It is possible to increase the reliability of the DT by modifying the model (5), for example, by using the LASSO method [18, 19]. The method is concerned with identifying the constraints of norm of a vector of coefficients of the model, which will lead to zero of some of its coefficients, i. e., in fact, the exclusion of one or more stations from expression (5). Also, an important positive effect arising from the use of the LASSO method is an increase in the stability and interpretability of the model, since, as a result, the features that have the greatest influence on the response vector are selected. In other word, at a zero value of the regularization parameter λ, the LASSO regression is reduced to the least squares (LS) method, and with its increase, the formed model becomes more and more "laconic" until it degenerates into the so-called null model, which gives the same output for all possible inputs [20]. This can be seen from the expression

$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta}\left(\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 + \lambda|\beta|\right), \quad (6)$$

where y is an expected model response.

At $\lambda = 1$, it is possible to reduce expression (5) to 3 terms (β_3, β_9, $\beta_{12} = 0$), thereby increasing the probability of the model triggering to 86.3%, while practically without losing accuracy (MSE ~ 12 nT) and maintaining the correlation parameters and the statistical homogeneity of the original and synthesized samples at the model level (5). It is even more significant to increase the probability of the model triggering, possibly excluding the maximum number of terms from expression (5), while controlling the constancy of the correlation parameter and the Student's t-test results, as well as keeping the MSE in some acceptable range, for example, MSE ≤ 30 nT.

However, according to previous experience, the implementation of this operation by simply increasing the parameter λ is ineffective and leads to a significant increase in the simulation error with a relatively small decrease in the number of its terms. In other words, further application of machine optimization methods (including ridge regression and Elastic Net [21]) is impractical, and the subsequent minimization of the number of features should be done manually, for example, by pairwise comparative analysis of the statistics of available predicates. For this purpose, we exclude the baseline from the time series of each station, normalize the histogram and on the basis of by Kolmogorov — Smirnov criteria select for the obtained samples $|\Delta X|$ the function that best approximates the distribution of its values. The function, in turn, in addition to the homogeneity of general samples, may indicate the homogeneity of the physical mechanisms responsible for the appearance of disturbances at the points of their observation [16]:

$$\left|\Delta X_{ij}\right| = \left|X_{ij} - \text{Me}\left(X_j\right)\right|, \quad (7)$$

where X_{ij} is the i-th value for j-th day of X-component at the station; Me(X_j) is a sample median X for j-th day; i and j correspond to the ordinal numbers of a minute in a day (from 1 to 1440) and a day in a year (from 1 to 365), respectively.

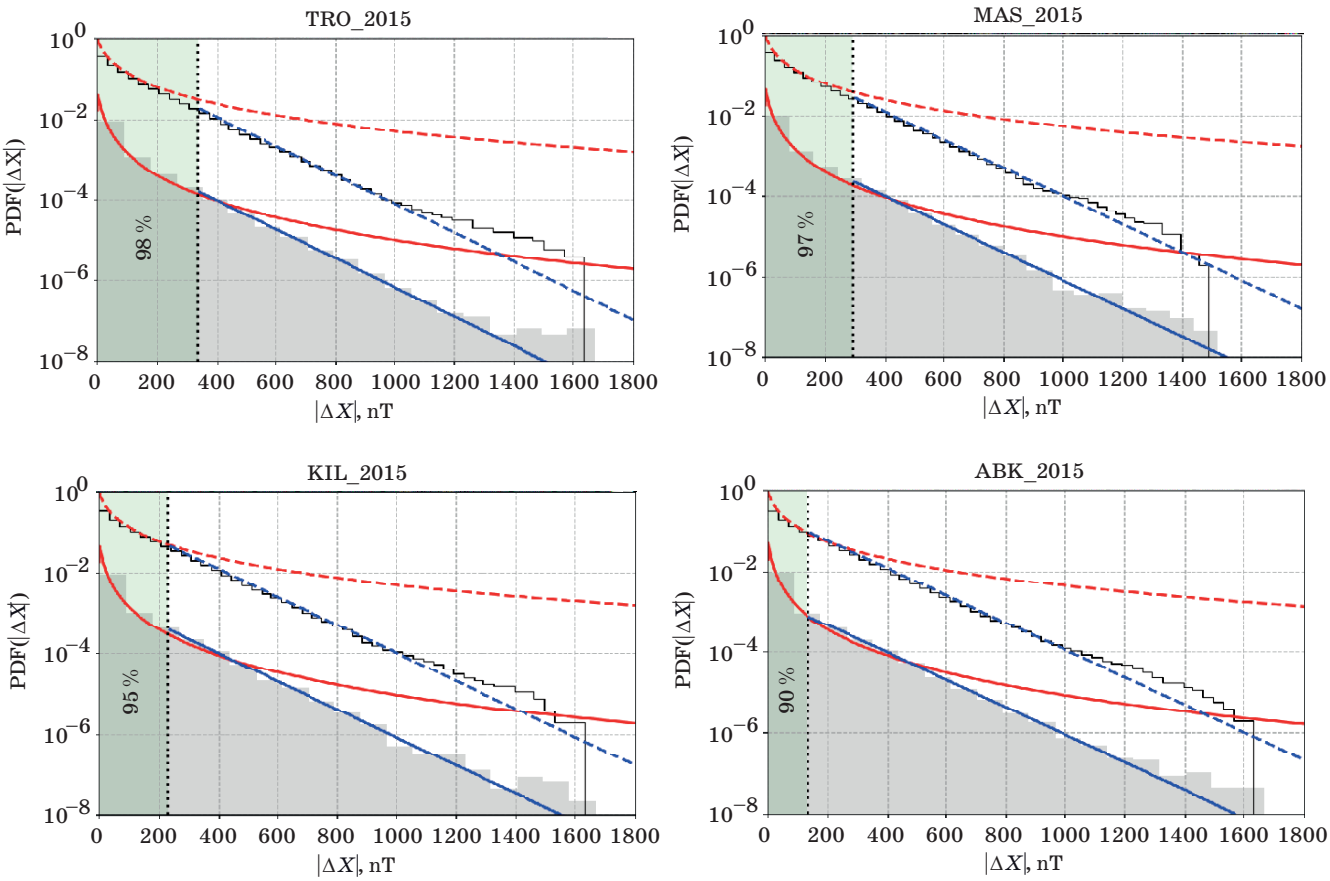
Analysis of the disturbed (i. e., in this case, excluding the daily variations of the GMF) X-components of the GMF at the KIL station ($|\Delta X|_{\text{KIL}}$) absolute values distribution demonstrated that most of the sample values are distributed according to the lognormal law (Fig. 1). However, starting from the 95th percentile, an exponential tail is observed, indicating that the variance of the studied value is determined mainly by rare intense (rather than frequent small) deviations, apparently in this case due to substorm activity. Further research demonstrated that the samples statistically closest to $|\Delta X|_{\text{KIL}}$ are $|\Delta X|_{\text{TRO}}$, $|\Delta X|_{\text{MAS}}$ and $|\Delta X|_{\text{ABK}}$, which are the absolute values of the disturbed components of the GMF X-component at stations Tromsø (TRO), Masi (MAS) and Abisko (ABK). In this case, almost the only difference is the sample percentile corresponding to the beginning of the exponential tail, which is apparently determined by the latitudinal location of a particular station (see Fig. 1, Table 1).

In addition, analysis of correlation between the regional IL-index (the intensity of the western auroral electrojet, i. e., the horizontal current flowing in the auroral region of the ionosphere) and the X-component of the four stations identified (see Fig. 1) revealed the proportionality of these correlations (in each case, the Pearson correlation coefficient is ~0.7), which again indicates that the stations under consideration are equally affected by the same external factors. Thus, datasets including data of TRO, MAS and ABK stations, are best suited for modeling the desired parameter. In this case, obviously, the minimum set of data sources can only consist of these stations. Taking this into account, expression (5) can be reduced to the following:

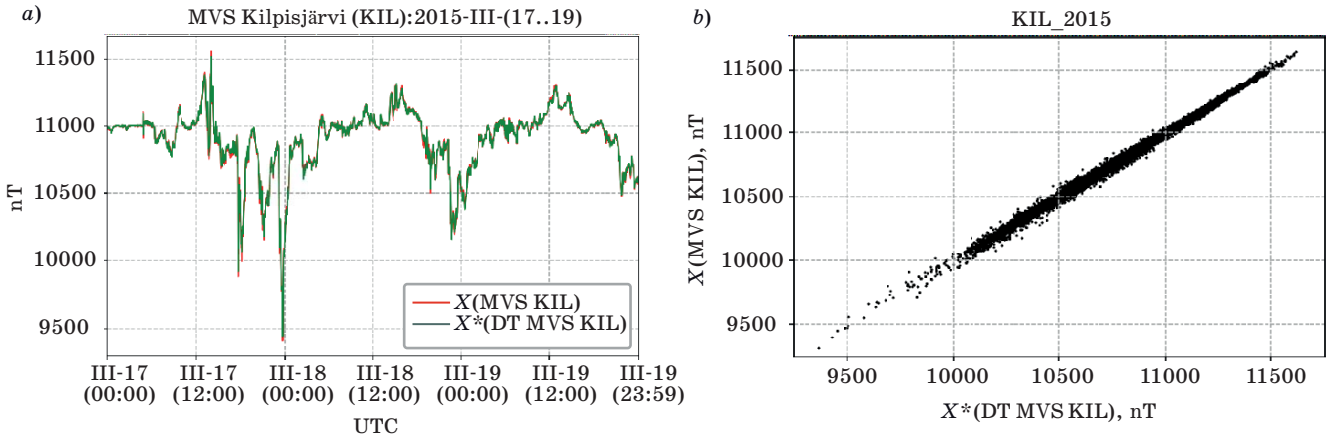
$$X^*_{\text{KIL}} = \alpha + \beta_4 X_{\text{NOR}} + \beta_5 X_{\text{MAS}} + \beta_8 X_{\text{ABK}}, \quad (8)$$

where $\alpha = 248.719$ nT; $\beta_4 = 0.2914795$; $\beta_5 = 0.286204$; $\beta_8 = 0.4405047$.

Figure 2, *a* represents the magnetograms of the initial time series and time series reconstructed on the basis of the regression model (8), which includes one of the most powerful magnetic storms over the past few years of observations. The dispersion of the simulation results can be estimated from the



■ ***Fig. 1.*** Statistics of the disturbed geomagnetic variations: red and blue solid (dashed) lines correspond to the probability density functions (survival) of the lognormal and exponential distribution laws, respectively; black solid line — empirical survival function; PDF — probability density function



■ ***Fig. 2.*** Verification of the digital twin of the station KIL: *a* — magnetograms of the initial time series; *b* — magnetograms of the time series reconstructed on the basis of the regression model

scattering diagram is demonstrated in Fig. 2, *b*. The probability of triggering a DT based on model (8) is 99.5%, and MSE < 30 nT (Table 3).

It should be noted that methods based on geospatial interpolation may be a possible alternative, and in some situations the only approach to creating a DT. For example, according to the Inverse Distance Weighting (IDW) method [22], the interpolated value of the parameter at a given geographical point is determined by the weighted average sum of deterministic values in its vicinity. In the case of Shepard's modification [22], the level of influence of the deterministic point on the desired value is set by the exponent p and with distance from the top of

■ ***Table 3.*** KIL station digital twin model validation parameters

Model	MSE, nT	MSE, %	r	Student's t-test		T_W, min	T_F, min	P_W, %
				Statistic	p-value			
Expr. (5), LS	11.5	0.51	0.999	~0	~1	406936	118664	77.423
Expr. (5), LASSO	12.0	0.54	0.999	~0	~1	453819	71781	86.343
Expr. (8), LASSO	28.9	1.25	0.999	~0	~1	523257	2343	99.554
Expr. (9), IDW ($p = 3$)	114.1	4.94	0.995	~0	~1	406936	118664	77.423

Note: P_W is the expected probability of the model being triggered.

the polygon, including the reference data sources, its influence on the interpolated value weakens. For the case under consideration, the ratio of the IDW method is as follows:

$$X^*_{\text{KIL}} = \sum_{i=1}^{m} \frac{1}{d_i^p} X_i \Big/ \sum_{i=1}^{m} \frac{1}{d_i^p}, \qquad (9)$$

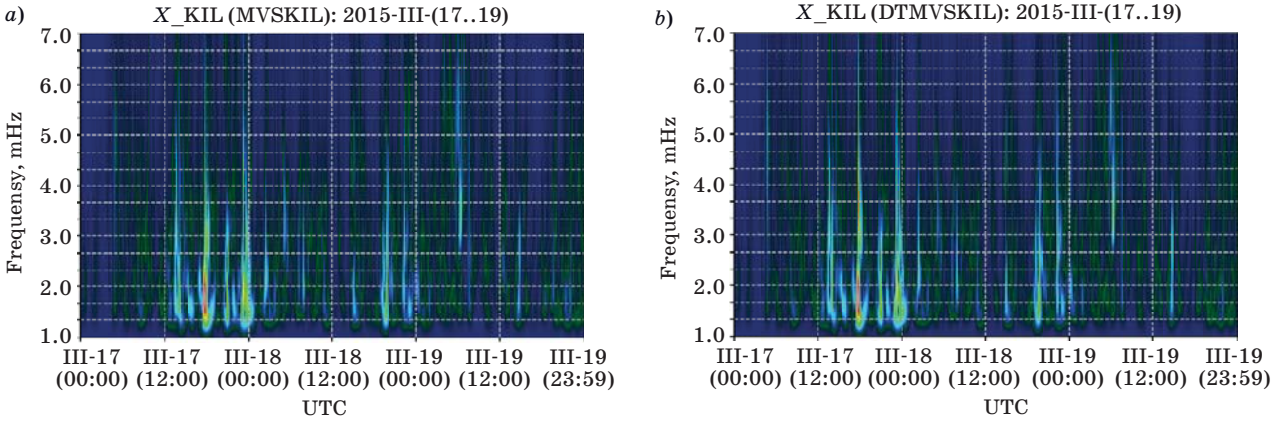
where m is a number of stations in the auroral cluster; d is a distance between the KIL station and the i-th station of the auroral cluster; p is a weight coefficient; X_i is a value of X-component of i-th station.

The disadvantage of the IDW method for interpolating geomagnetic disturbances is the assumption that the disturbance field is isotropic in it. However, here it should be taken into account that latitudinal and longitudinal scales of most geomagnetic disturbances differ significantly. Research results have shown that in relation to the problem under consideration, the MSE of the DT model built on the basis of the IDW method monotonically increases with decreasing p, which indicates that the sought parameter is determined mainly by the data of the stations closest to the modeled object. As a result, the modeling error by means of expression (9) will be slightly higher than the MSE of the regression models (see Table 3). However, despite this, the geospatial interpolation method can be useful in the absence of a response vector, i. e., in the situation when there is no physical prototype of the station.

Digital twin verification in frequency domain

Although variations in the GMF in the range of periods of 2–12 min significantly inferior in intensity to global geomagnetic disturbances — magnetic storms and substorms — they are still extremely important.

Disturbances in this frequency range (Pi3 / Ps6 pulsations, Pc5 waves, the beginnings of substorms) lead to the most powerful bursts of geoinduced currents in power lines. Therefore, an important aspect in the functioning of the DT is the identification and storage of information about these disturbances. Let us select by means of the Butterworth high-pass filter in the X_{KIL} and X^*_{KIL}



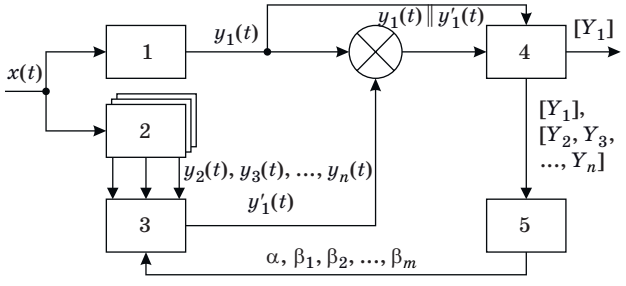
■ ***Fig. 3.*** Verification of the digital twin of the magnetic station KIL in the frequency range of 1–7 mHz: wavelet scalogram of original (*a*) and recovered (*b*) time series

Thus, from Fig. 3, *a* and *b* as well as from a number of similar tests for other fragments of the time series, it follows that in the region of ultra-low frequencies (with periods of 2–12 min), insignificant (within the limits of the error stated in Table 3) deviations of the amplitude are observed, while the spatial localization of frequency packets remains practically unchanged.

Integration of the digital twin into the process of collecting geomagnetic data

Figure 4 schematically demonstrates the model of integration of the DT of magnetic station into the processes of collecting and registering geomagnetic data. So, according to the proposed scheme, the disturbing effect $x(t)$ extends to the physical prototype of the magnetic station (1) and a number of reference data sources (2), involved in the base of the DT models (3).

Depending on the number m of stations available at the time t_i, a model that provides the minimum error is selected, by means of which the DT of the magnetic station (1) generates the corresponding value $y'_1(t_i)$. Further, the data corresponding to the state of the GMF at the i-th moment of time, from the output of the DT and its physical prototype, are sent to the comparison device, which, by comparing these values, makes a decision on registration as a measurement result or data from a magnetic station, for example, based on the fulfillment of the condition (10), or its DT (in cases of its failure), while the value of the magnetic station is also saved, however, it is marked as anomalous. If there is no output signal from the magnetic station, then the DC value is recorded as the measurement result. The verified values stored in the geomagnetic database (4) are structured in the form of response vectors and regressors and are used to update and adjust the vectors of coefficients of the DT models (5).



■ ***Fig. 4.*** Model of digital twin integration into the processes of collection and registration of geomagnetic data: *1* — magnetic station; *2* — reference magnetic stations; *3* — digital twin of the magnetic station; *4* — data base; *5* — machine learning system

$$\left|x_i - x_i^*\right| < 3\sigma \quad \text{or}$$

$$\left|x_i - x_i^*\right| < 3\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\left(x_i - x_i^*\right) - \bar{x}\right)^2}, \qquad (10)$$

where σ is a standard deviation; x_i^* and x_i are the values of the digital twin and its physical prototype, respectively, at the i-th moment of time t.

Figure 5 on the example of the KIL station demonstrates an algorithm that explains the diagram shown in Fig. 4.

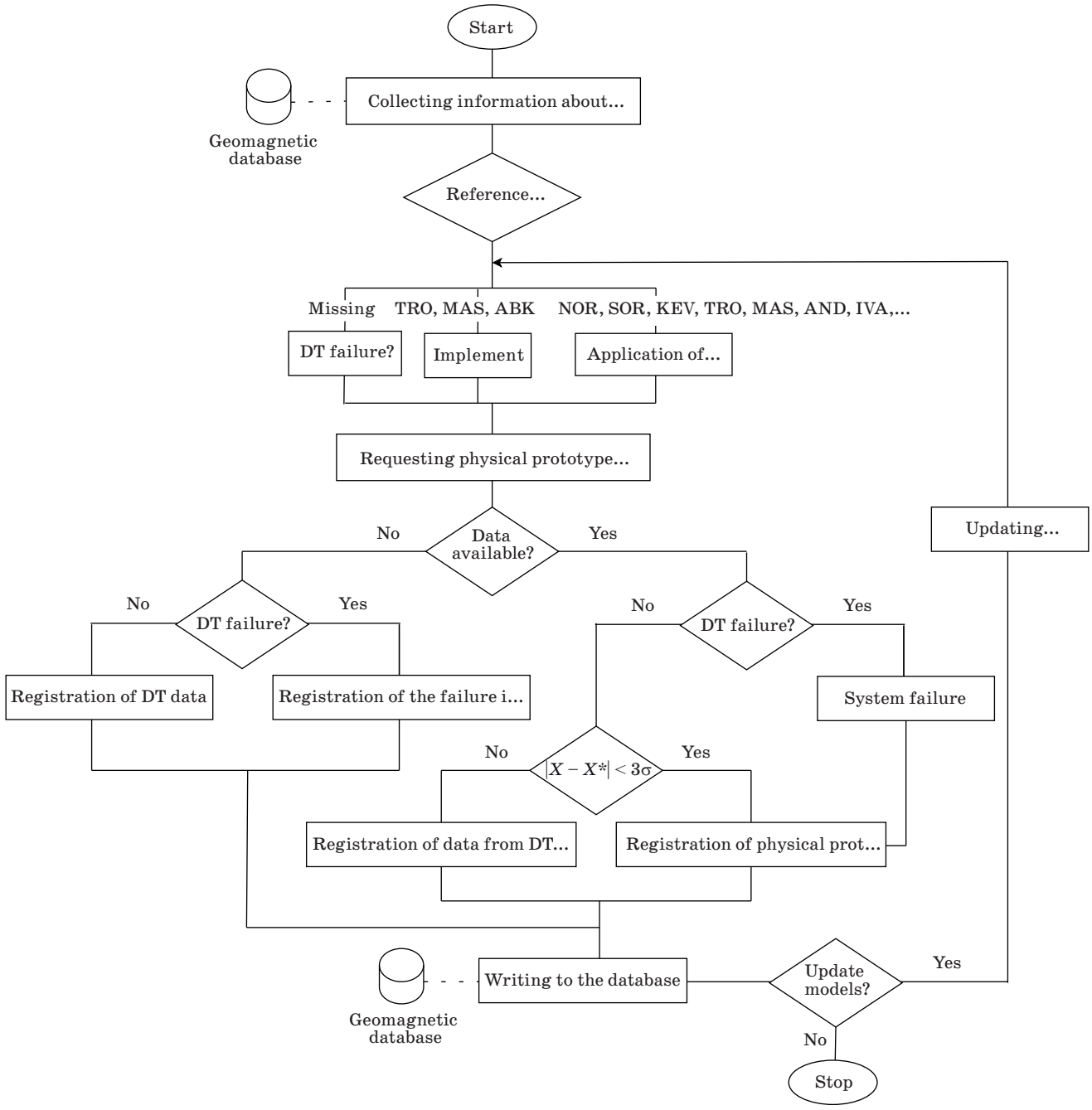
Thus, the application of the proposed scheme and algorithm in the case of the KIL station makes it possible to recover 99.55% of the data for 2015, while the MSE of 86.73% of the recovered values does not exceed 12 nT. As follows from the algorithm (see Fig. 5), the state of failure of the entire local system for collecting and registering geomagnetic data occurs with the simultaneous absence of a signal at the output of the magnetic station and its DT. For the KIL station, the calculated value of the probability of such an event occurring is less than 0.0016%, which corresponds to eight missing values per year, which, in turn, can be restored using linear interpolation methods.

Discussion of the results and prospects for their application

As has been shown, the introduction of magnetic station DT into the processes of collecting and registering geomagnetic data due to the redundancy effect can (at the data consumer level) significantly increase the reliability and fault tolerance of individual magnetometers, as well as reduce the labor intensity of preprocessing of geomagnetic data, for example, such as search and identification of outliers in time series.

However, when implementing the approach, it is necessary to take into account the limitations of its effective application, which are determined, first of all, by the spatial anisotropy of the GMF parameters. Thus, the MSE of the DT for each specific case (magnetic station) will differ, depending on the geographic location of this physical prototype, as well as the number and distance of the surrounding magnetic stations. At the same time, the general methodology for selecting reference stations, synthesis and optimization of regression models will practically not change.

A perspective in the development of virtual magnetic stations is the integration of GMF satellite observation data (for example, SWARM, CHAMP missions, etc.) into the information environment of the DT. It can be assumed that the implementation of the approach, in addition to the aggregation of ad-



■ ***Fig. 5.*** Algorithm of the process of geomagnetic data collecting and registering with the implementation of the digital twin on the example of the KIL magnetic station

ditional data required for the calibration (settings of models) of the DT of magnetic stations, can also weaken a number of methodological limitations of the effective use of the DTs, associated, for example, with the absence of nearby magnetic stations.

Speaking about the prospects of using the DT of magnetic stations, the following tasks should mainly be highlighted:

— reconstruction of geomagnetic data time series;

— automated search and identification of outliers in geomagnetic data time series;

— collection of geomagnetic data in conditions where the use of physical magnetic stations is unacceptable or ineffective, for example, in the immediate vicinity of objects that have a strong noisy effect on magnetic sensors and primary measuring transducers (trunk pipelines, power lines, railway and oil and gas infrastructure facilities, etc.).

— information support of the processes of directional drilling of deep wells in the Arctic zone of the Russian Federation [23, 24].

Also, it should be noted here that DTs have the potential to be used in problems of machine search and identification of localized GMF disturbances, for example, such as MPE (magnetic perturbation events), which are isolated bursts of field intensity with a duration of 5–15 min at night [25] and can be responsible for intense bursts of geoinduced currents in power lines [26]. The horizontal scale of such disturbances is ~200–300 km, and they are recorded, as a rule, at 1–2 stations of the network. Thus, DTs are able to automate this process by isolating disturbances that sharply differ from the model values.

Conclusion

In this paper (using the KIL magnetic station as an example), it is shown that the DTs of magnetic stations built on the basis of LASSO regression are capable of providing retrospective forecast and restoration of the *X*-component of the GMF vector in the auroral zone with a mean square error from 11.5 (in 77.4% of cases) to 29 nT (in 99.6% of cases) depending on the number of reference stations used.

Comparative analysis of wavelet spectrograms of data from the magnetic station DT and its physical prototype in the frequency range with periods of 2–12 min (Pi3 / Ps6 pulsations, Pc5 waves, the onset of substorms) showed that in the amplitude region of the information signal there may be minor differences commensurate with modeling error, however, the spatial localization of frequency packets remains practically unchanged.

In the absence of a physical prototype of the magnetic station (the response vector of the training sample), the implementation of the DT is possible on the basis of spatial interpolation methods, but here one should expect a slightly larger (compared to the regression approach) modeling error.

The main factors limiting the effectiveness of the proposed approach are the specifics of the geographic location of a particular physical prototype, as well as the number and distance of nearby magnetic stations. It is possible to minimize the influence of these factors by expanding the information environment of the DT, for example, by aggregating data from satellite observations of the GMF.

Financial support

This work was supported by a grant from the Russian Science Foundation No. 21-77-30010, and also partially supported by grants from the Russian Foundation for Basic Research No. 20-07-00011-a and the Expert Center “Project Office for the Development of the Arctic” (Agreement No. 217-G dated January 13, 2021).

Acknowledgements

We thank the institutes who maintain the IMAGE Magnetometer Array: Tromsø Geophysical Observatory of UiT the Arctic University of Norway (Norway), Finnish Meteorological Institute (Finland), Institute of Geophysics Polish Academy of Sciences (Poland), GFZ German Research Centre for Geosciences (Germany), Geological Survey of Sweden (Sweden), Swedish Institute of Space Physics (Sweden), Sodankylä Geophysical Observatory of the University of Oulu (Finland), and Polar Geophysical Institute (Russia).

References

1. Love J. An international network of magnetic observatories. *EOS, Transactions, American Geophysical Union*, 2013, vol. 94, no. 42, pp. 373–384.
2. Khomutov S. Yu. International project INTERMAGNET and magnetic observatories of Russia: cooperation and progress. *E3S Web of Conferences*, 2018, no. 62, pp. 02008. doi:10.1051/e3sconf/2018620
3. Vorobev A. V., Vorobeva G. R. Approach to assessment of the relative informational efficiency of INTERMAGNET magnetic observatories. *Geomagnetism and Aeronomy*, 2018, vol. 58, no. 5, pp. 648–652 (In Russian). doi:10.1134/S0016793218050158
4. Vorobev A. V., Pilipenko V. A., Enikeev T. A., Vorobeva G. R. Geoinformation system for analyzing the dynamics of extreme geomagnetic disturbances from observations of ground stations. *Computer Optics*, 2020, vol. 44, no. 5, pp. 782–790 (In Russian). doi:10.18287/2412-6179-CO-707
5. Reich K., Roussanova E. Visualising geomagnetic data by means of corresponding observations. *Int J Geomath*, 2013, no. 4, pp. 1–25. doi:10.1007/s13137-012-0043-4
6. Gvishiani A. D., Lukianova R. Yu., Soloviev A. A. *Geomagnetizm: ot yadra Zemli do Solnca* [Geomagnetism: from the core of the Earth to the Sun]. Moscow, Rossijskaya akademiya nauk Publ., 2019. 186 p. (In Russian).
7. Gvishiani A. D., Agayan S. M., Bogoutdinov Sh. R., Kagan A. I. Gravitational smoothing of time series. *Trudy Instituta matematiki i mekhaniki UrO RAN*, 2011, vol. 17, no. 2, pp. 62–70 (In Russian).
8. Mandrikova O. V., Soloviev I. S. Wavelet technology for processing and analyzing geomagnetic data. *Ci-*

frovaya obrabotka signalov, 2012, no. 2, pp. 24–28 (In Russian).
9. Mandrikova O. V., Solovyev I. S., Khomutov S. Y., Geppener V. V., Klionskiy D. M., Bogachev M. I. Multiscale variation model and activity level estimation algorithm of the Earth's magnetic field based on wavelet packets. *Ann. Geophys.*, 2018, no. 36, pp. 1207–1225. doi:10.5194/angeo-36-1207-2018
10. Kondrashov D., Shprits Y., Ghil M. Gap filling of solar wind data by singular spectrum analysis. *Geophys. Res. Lett.*, 2010, vol. 37, L15101, doi:10.1029/2010GL044138
11. Vorobev A. V., Vorobeva G. R. Inductive method of geomagnetic data time series recovering. *SPIIRAS Proceedings*, 2018, no. 2, pp. 104–133 (In Russian). doi:10.15622/sp.57.5
12. Parmar R., Leiponen A., Llewellyn D. W. T. Building an organizational digital twin. *Business Horizons*, 2020, vol. 63, no. 6, pp. 725–736. doi:10.1016/j.bushor.2020.08.001
13. Zongyan W. *Digital Twin Technology*. In: *Industry 4.0 — Impact on Intelligent Logistics and Manufacturing*. IntechOpen. Pp. 95–114. doi:10.5772/intechopen.80974
14. Tanskanen E. I. A comprehensive high-throughput analysis of substorms observed by IMAGE magnetometer network: Years 1993–2003 examined. *J. Geophys. Res.*, 2009, no. 114, p. A05204. doi:10.1029/2008JA013682
15. Demyanov V. V., Savelyeva E. A. *Geostatistika. Teoriya i praktika* [Geostatistics. Theory and practice]. Moscow, Nauka Publ., 2010. 327 p. (In Russian).
16. Vorobev A., Vorobeva G. *Properties and type of latitudinal dependence of statistical distribution of geomagnetic field variations*. In: *Kocharyan G., Lyakhov A. (eds). Trigger Effects in Geosystems. Springer Proceedings in Earth and Environmental Sciences*. Springer, Cham., 2019. Pp. 187–196. https://doi.org/10.1007/978-3-030-31970-0_22
17. Vorobev A. V., Vorobeva G. R. Correlation analysis of geomagnetic data synchronously recorded by the INTERMAGNET magnetic laboratories. *Geomagnetism and Aeronomy*, 2018, vol. 58, no. 2, pp. 187–193 (In Russian). doi:10.1134/S0016793218020196
18. She Yiyuan. Sparse regression with exact clustering. *Electron. J. Statist.*, 2010, vol. 4, pp. 1055–1096. doi:10.1214/10-EJS578
19. Hoerl R. W. Ridge regression: a historical context. *Technometrics*, 2020, vol. 62, no. 4, pp. 420–425. doi:10.1080/00401706.2020.1742207
20. Tokmakova A. A., Strijov V. V. Estimation of linear model hyperparameters for noise or correlated feature selection problem. *Informatics and Applications*, 2012, vol. 6, no. 4, pp. 66–75 (In Russian).
21. Zou H., Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2005, no. 67, pp. 301–320. doi:10.1111/j.1467-9868.2005.00503.x
22. Isaaks E. H., Mohan R. *An Introduction to Applied Geostatistics*. Oxford, Oxford University Press, 1989. 592 p.
23. Gvishiani A. D., Lukianova R. Yu. Study of the geomagnetic field and problems of accuracy of directional drilling in the Arctic region. *Izvestiya vysshikh uchebnykh zavedenii. Gornyi zhurnal*, 2015, no. 10, pp. 94–99 (In Russian). doi:10.17580/gzh.2015.10.17
24. Gvishiani A. D., Lukianova R. Yu. Estimating the influence of geomagnetic disturbances on the trajectory of the directional drilling of deep wells in the Arctic region. *Izvestiya. Physics of the Solid Earth*, 2018, no. 4, pp. 19–30 (In Russian). doi:10.1134/S0002333718040051
25. Engebretson M. J., Steinmetz E. S., Posch J. L., et al. Nighttime magnetic perturbation events observed in Arctic Canada: 2. Multiple-instrument observations. *Journal of Geophysical Research: Space Physics*, 2019, no. 124, pp. 7459–7476. https://doi.org/10.1029/2019JA026797
26. Datcu M., Le Moigne J., Loekken S., Soille P., Xia G.-S. Special issue on big data from space. *IEEE Transactions on Big Data*, 2020, vol. 6, no. 3, pp. 427–429. doi:10.1109/TBDATA.2020.3015536

УДК 004.94
doi:10.31799/1684-8853-2021-2-60-71

Методология создания и перспективы применения проблемно-ориентированных цифровых двойников магнитных обсерваторий и вариационных станций

А. В. Воробьев[а,б], канд. техн. наук, доцент, orcid.org/0000-0002-9680-5609, geomagnet@list.ru
В. А. Пилипенко[б,в], доктор физ.-мат. наук, главный научный сотрудник, orcid.org/0000-0003-3056-7465
Г. Р. Воробьева[а], канд. техн. наук, доцент, orcid.org/0000-0001-7878-9724
О. И. Христодуло[а], доктор техн. наук, доцент, orcid.org/0000-0002-3987-6582
[а]Уфимский государственный авиационный технический университет, К. Маркса ул., 12, Уфа, 450008, РФ
[б]Геофизический центр РАН, Молодежная ул., 3, Москва, 119296, РФ
[в]Институт физики Земли им. О. Ю. Шмидта РАН, Б. Грузинская ул., 10, стр. 1, Москва, 123995, РФ

Введение: магнитные станции являются одним из основных инструментов наблюдения геомагнитного поля, однако пропуски и аномалии во временных рядах геомагнитных данных, нередко превышающие 30 % от числа зарегистрированных значений, негативно отражаются на эффективности реализуемого подхода и затрудняют применение элементов математического обеспечения, требующих соблюдения условия непрерывности информационного сигнала. Кроме этого, отсутствующие значения вносят дополнительную неопределенность в задачах компьютерного моделирования динамики пространственного распределения параметров геомагнитных вариаций. **Цель:** разработать методологию повышения эффективности технических средств наблюдения геомагнитного поля. **Метод:** создание и интеграция в процессы сбора и предварительной обработки геомагнитных данных проблемно-ориентированных цифровых двойников магнитных станций, позволяющих с известной точностью имитировать функционирование их физических прототипов. **Результаты:** на примере магнитной станции Kilpisjärvi (Финляндия) показано, что использование цифровых двойников, информационную среду которых составляют геомагнитные данные окрестных станций, позволяет провести восстановление (ретроспективный прогноз) параметров геомагнитных вариаций со среднеквадратической ошибкой в авроральной зоне до 11,5 нТл. Интеграция проблемно-ориентированных цифровых двойников магнитных станций в процессы сбора и регистрации геомагнитных данных способна обеспечить автоматическую идентификацию и замещение отсутствующих и аномальных значений, повышая за счет эффекта резервирования отказоустойчивость магнитной станции как объекта-источника данных. Так, например, цифровой двойник станции Kilpisjärvi реализует восстановление 99,55 % годовой информации, из них 86,73 % с ошибкой, не превышающей 12 нТл. **Обсуждение:** по причине пространственной анизотропии параметров геомагнитного поля ошибка на выходе цифрового двойника для каждого конкретного случая будет отличаться в зависимости от географического местоположения магнитной станции, а также числа и удаленности окрестных магнитных станций. Однако данную проблему возможно минимизировать, интегрируя в информационную среду цифрового двойника геомагнитные данные спутниковых наблюдений. **Практическая значимость:** применение предложенной методологии делает возможными автоматизированную диагностику временных рядов геомагнитных данных на предмет выбросов и аномалий, а также восстановление отсутствующих значений и идентификацию мелкомасштабных возмущений.

Ключевые слова — цифровые двойники, восстановление временных рядов, статистический анализ, геомагнитные данные, магнитные станции.

Для цитирования: Vorobev A. V., Pilipenko V. A., Vorobeva G. R., Khristodulo O. I. Development and application of problem-oriented digital twins for magnetic observatories and variation stations. *Информационно-управляющие системы*, 2021, № 2, с. 60–71. doi:10.31799/1684-8853-2021-2-60-71

For citation: Vorobev A. V., Pilipenko V. A., Vorobeva G. R., Khristodulo O. I. Development and application of problem-oriented digital twins for magnetic observatories and variation stations. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2021, no. 2, pp. 60–71. doi:10.31799/1684-8853-2021-2-60-71